

Decision Science

INTERNAL ASSIGNMENT- JUNE 2020



**NMIMS GLOBAL ACCESS
SCHOOL FOR
CONTINUING EDUCATION**

**NMIMS Global Access
School for Continuing Education (NGA-SCE)**

UNDER THE GUIDANCE OF

Prof. Dr. N. Palaniappan

SUBMITTED BY:

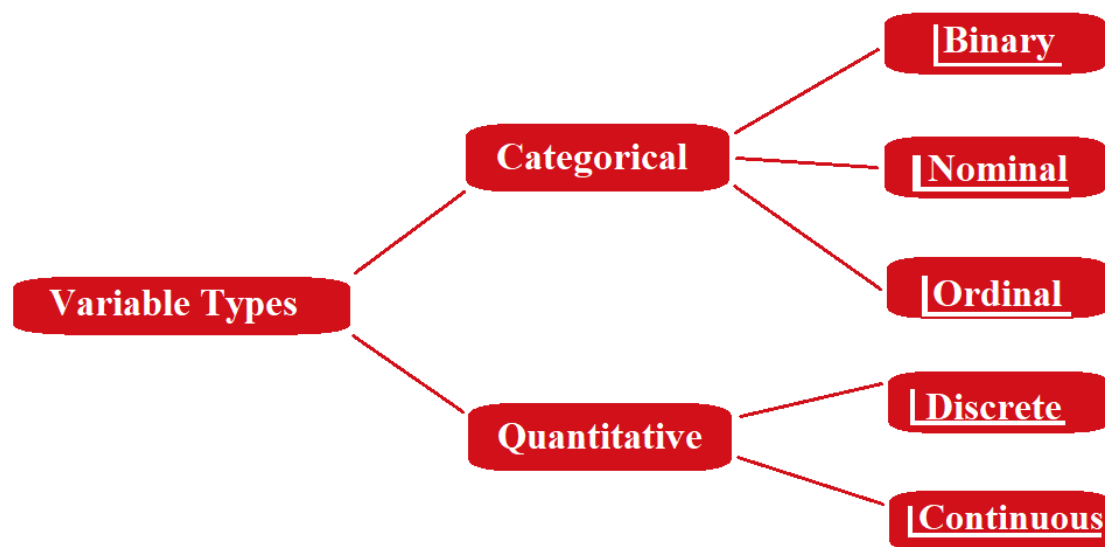
**DIKSHA GUPTA
M.B.A. 2ND YEAR**

**DATE OF SUBMISSION:
June 23, 2020.**

1. Identify the types of the variable.

In Decision Science or Statistics, a 'variable' can be defined as an attribute to be studied, of any object. Selecting, which variable attributes to measure can be a good design for experiments.

Types of data: Quantitative and Categorical variables



Data contains a specific measurement of the set of variables. These variables are generally divided into 2 types:

- **Quantitative variables:**

Quantitative data represents quantities or amounts. While collecting quantitative data, the recorded numbers represent operations with arithmetic such as add, subtract, divide etc.

Quantitative variable types	Representation by the data	Examples
Discrete (integer) variables	Counting individual items, values or parameters.	→ Number of trees. → Number of students in a class.
Continuous (floating) variables	Measurements of non- integral or continuous values.	→ Field intensity. → Distance. → Weight or BMI. → Salary.

- **Categorical variables:**

Categorical data represents groups.

Categorical variable types	Representation by the data	Examples
Binary (dichotomous variables)	Yes/ No or Present/ absent etc.	→ Heads/ tails in a coin toss. → Whether loan availed.
Nominal variables	Groups with no rank/ order among them.	→ Colours. → Brands.
Ordinal variables	Groups that are ranked in a specific order, but not representing quantities.	→ Hypotension/ normal/ hypertension. → Material quality*. → Survey's rating scale response from 1 to 10.

*Sometimes a nominal variable can also be treated as a quantitative variable. A variable may fall under more than one variable types. If the scale is numeric, such a material quality or survey ratings, despite having floating values, the scale can fall under continuous quantitative data- type.

Sr. No.	Variable	Data Type	Values
a	Gender	Nominal Categorical variable	Male, Female and Others.
b	Educational background	Ordinal Categorical variable	None(0), Metric pass(12), Graduate(15), Post- graduate(17), Doctorate(21) etc.
c	Satisfaction	Ordinal Categorical variable	Low, Medium, High or measured on a 1 to 5 level scale.
d	Motivation	Ordinal Categorical variable	Low, Medium, High or measured on a 1 to 5 level scale.
e	Exchange rate	Continuous Quantitative variable	Any real number value such as 0.5, 78 etc.
f	Gold price	Continuous Quantitative variable	Any real number value such as 0.5, 78 etc.
g	Preference of cars	Nominal Categorical variable	SUV, Sedan or Race car.
h	Teacher's feedback	Ordinal Categorical variable	Bad(1), Good(2), Extraordinary(3) etc.
i	Grades in Post- Graduation	Ordinal Categorical variable	Grade O, Grade A, Grade B, Grade F etc.
j	Marital status	Nominal Categorical variable	Married, Unmarried, Divorced, Widow/Widower.
k	Quality of services	Ordinal Categorical variable	Bad(1), Good(2), Extraordinary(3) etc.

l	Age group	Ordinal Categorical variable	Young (0-12), Teenage (13-19), Adult (20-50), Senior (beyond 50).
m	GDP	Continuous Quantitative variable	Any real numbers such as \$3000 billion.
n	Interest rate	Continuous Quantitative variable	Any real numbers such as 10%, 5.8%, -1.3% etc.
o	Twitter comments	Nominal Categorical variable	Exclamatory, Liked, Positive, Disliked, Joyful etc.
p	Facebook pictures	Nominal Categorical variable or Discrete Quantitative variable.	No ordered ranking among the images. OR Images with pixel values from 0 to 255 (integers).

a. Gender: (Nominal)

When gender has just two categories simply, named as Male and Female, the gender data can be treated as a binary variable. It can also be converted into discrete variable by writing Male as 1 and Female as 0 or vice-versa. The word binary means, of relating to two (bi).

When gender has more than two categories such as Male, Female and others, the gender can be treated as a nominal variable. There is no ordered ranking among the three categories.

b. Educational background: (Ordinal)

Educational background can be treated as the variable, having the highest level of education accomplished by someone. Furthermore, these categories can be ranked in an ordered manner or level. Here, one can give ordered level or ranking.

c. Satisfaction: (Ordinal)

The level of satisfaction cannot be counted in numbers, as it is not a real object. However, one can grade it between 1 and 5 and compare it with others. Hence, it can be treated as an ordinal variable.

d. Motivation: (Ordinal)

Similar to the level of satisfaction one is having, the amount of impact one has gone through, by encountering some motivations, can be compared with the others.

The impact on one's life, due to some motivational speeches can be scaled as Low, Medium or High.

e. Exchange rate: (Continuous)

\$1 = ₹75 or ₹1 = \$0.013.

The rate of exchange can acquire any real value. Above example illustrates the exchange rate to be 75 or 0.013 (real). Furthermore, there is sometimes a possibility of negative rates also.

f. Gold price: (Continuous)

Similar to the exchange rate, gold prices also can be any real values such as 80₹/gm. Few days ago, crude oil prices had gone negative in USA. Similar maybe the cases rarely, for golds as well.

g. Preference of cars: (Nominal)

A person with huge family might prefer Sedan or SUVs and a rich person might prefer a race car such as Ferrari. These preferences cannot be explained in terms of orders rankings.

h. Teacher's feedback: (Ordinal)

If a feedback is represented in terms of overall ratings, the feedback can be ordered. Good feedback can be ranked higher the bad one.

i. Grades in Post- Graduation: (Ordinal)

Grade O, Grade A can be ranked in order. Grade A is better than B, hence can be given higher ranking. However, it is different from 0.0 – 100.0 (percentage) and 1.0 – 8.0 (CGPA), that would have been continuous variables.

j. Marital status: (Nominal)

The status shouldn't be ranked in any order. Ordered representation for married and widower persons cannot be possible.

k. Quality of services: (Ordinal)

Similar to the teacher's feedback, the quality can be measured in ordered manner. A good quality can be ranked higher than the bad one.

l. Age group: (Ordinal)

Youngsters, adults and senior citizens can be arranged in an ordered manner, acquiring categorical values.

m. GDP: (Continuous)

GDP of India was 2718.73 and 2800 billion USD, respectively in 2018 and 2019. The values can be any real numbers, depending upon economical progress and several other factors.

n. Interest rate: (Continuous)

Japanese banks have negative interest rates, for current accounts. Whereas, Indian banks have rates around 5.0%. The values can hold any real number.

o. Twitter comments: (Nominal)

Comments can be in support or ranging protests. They might be joyful or sad. There is no ordered ranking.

p. Facebook pictures:

(Nominal)- If there can be no quantitative or ordered ranking among the Facebook pictures/ images, they can be kept classified with the unordered categorical data.

(Discrete)- A 1-bit monochrome image consists of pixel values as either 0 or 1 only, i.e. binary values. An 8-bit grayscale/ coloured image has pixel values from 0 to 255 = $2^8 - 1$, i.e. discrete or natural number values.

Example- This is a 1-bit monochrome image (binary). Here among the 2x2 image (four) pixels shown, white represents 1 and black colour represents 0.



2. Following data of performance scores is available of employees working with a company. You are required to perform the following:

a. Make the frequency distribution. Calculate the frequency and the Cumulative frequency.

Procedure to find Frequency:

- The 'frequency' (f) of a data value, is defined as the number of times the particular data value occurs/ repeats.
- For example, the performance score value 31 is repeated four times, in the given data set. Hence, the frequency of score 31 is 4. It can be represented as: $f_{31} = 4$.
- Similarly, the frequencies of each data value (performance scores ranging from 0 to 100) is to be counted.
- Note: Frequency, for the scores that are not observed in the dataset, is written as 0. For example, $f_{18} = 0$.

Procedure to find Cumulative Frequency:

- 'Cumulative frequency' (CF) is used to determine the number of observations that lie below a particular value in a data set.

$$CF_n = \sum f_{\leq n}.$$

$$CF_{48} = \sum f_{\leq 48} = f_0 + f_1 + f_2 + \dots + f_{48} \quad (\text{for discrete values of observations}).$$

- The cumulative frequency is calculated by adding every frequency from a frequency distribution table upto the sum of its predecessors.
- Cumulative frequency can also be calculated by the total of all the frequencies of all the observation upto a particular value.
- The last value of cumulative frequency will definitely be equal to the total number of all the observations.
- For example, $CF_{30} = f_0 + f_1 + \dots + f_{29} + f_{30} = 0+0+\dots+0+1 = 1$.
 $CF_{32} = f_{30} + f_{31} + f_{32} = 1+3+2 = 6$.

For the given data in the sets of 10 scores at a time, frequency and cumulative frequency are observed as:

Score Range	Frequency
0-29	0
30-39	36
40-49	27
50-59	32
60-69	33
70-79	21
80-89	26
90-100	33

Score Range	Cumulative Frequency
0-29	0
≤ 39	36
≤ 49	63
≤ 59	95
≤ 69	128
≤ 79	149
≤ 89	175
≤ 100	208

Observations:

- It is observed that the performance scores from 0 to 29, 99 and 100 are not obtained by any of the 208 employees.

$$f_0 = \dots = f_{29} = 0.$$

- According to the given data, 30 and 98 are the lowest and highest scores obtained by any employee, respectively.
- Additionally, the performance scores 65 and 81 are also not obtained by anybody. Hence, it can be said that:

$$f_{99} = f_{100} = f_{65} = f_{81} = 0.$$

- Performance score 53 is the most frequent score. $f_{53} = 8$.

=SUM(\$B\$2:T2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Unique Scores	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	
2	Frequency	1	3	2	6	6	2	6	1	4	5	5	1	4	3	2	3	2	
3	Cumulative Frequency	1	4	6	12	18	20	26	27	31	36	41	42	46	49	51	54	56	
4																			
5																			
6																			
7	Unique Scores	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
8	Frequency	2	4	1	2	3	1	8	5	4	2	2	2	3	3	5	4	5	4
9	Cumulative Frequency	58	62↑	63	65	68	69	77	82	86	88	90	92	95	98	103	107	112	116
10																			
11																			
12																			
13	Unique Scores	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	
14	Frequency	0	4	2	4	2	2	2	1	1	2	2	3	5	1	2	3	0	
15	Cumulative Frequency	116	120	122	126	128	130	132	133	134	136	138	141	146	147	149	152	152	
16																			
17																			
18																			
19	Unique Scores	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
20	Frequency	2	4	1	2	4	1	5	4	3	5	5	4	4	5	2	1	4	0
21	Cumulative Frequency	154	158	159	161	165	166	171	175	178	183	188	192	196	201	203	204	208	208

b. Calculate the mean, median, mode and quartiles.

Arithmetic Mean (Average):

- An 'arithmetic mean' (μ) is a single number, that summarisingly represents a list of numbers. It is the half way through (middle of) all the observed data set.

For example, the set {3,4,5,6,7} has the half way (average) at 5.

- It is a sum of the list of the observation values given and then divided by the total number of observations.

$$\text{Arithmetic Mean (A.M.)} = \frac{\text{Sum of all the observation values}}{\text{Number of observations}}$$
$$\mu = \frac{\sum x}{n}.$$

- In the given dataset, the number of employees (observations) is 208. Their performance scores (observation values) ranges from 30 to 98.

- According to the given data set:

$$\mu = \frac{52+57+50+68+74+\dots+66}{208} = 63.35096 .$$

Median:

- The 'median' is the middle number obtained in an ascending/ descending order sorted list.

It can be more descriptive of any data set than the average.

- The median is generally used, when there are outliers in the dataset, causing skewness. Hence, median tends to avoid the outliers and skewness in any data.

- For example, the set {5,2,6,3,7} can be ascendingly sorted as {2,3,5,6,7} and hence the median is 5.

- A median separates the higher half (50%tile) from the lower half of a data sample.

$$\text{Median} = \{(n + 1) \div 2\}^{\text{th}} \text{ value in a sorted dataset.}$$

- In the given dataset, sort dataset would be: {30,31,31,31,32,32,33,...,98}.

$$\text{Median} = 62.$$

Mode:

- The observation with highest frequency is known as the 'mode' of the data. It indicates, most of the data is crowded around which particular value.

- The mode of a set of data is the value in the set, that occurs most often.

- In the given dataset, it is observed that the performance score of 53 has been obtained by most of the people.

$$f_{\max} = f_{53} = 8.$$

$$\text{Mode} = 53.$$

- Note: For the given data, Mode < Mean and Median < Mean.

Hence, the data may be positively skewed, approximately. Several data-points may lie on the right side of the mode value. To know this, skewness is also to be measured.

Geometric Mean (G.M.):

- It is the central number in a G.P. (geometric progression), such as 3, 9, 27 has G.M. as 9.

$$G.M. = (\pi x_i)^{1/n} = (x_1 \cdot x_2 \dots x_n)^{1/n}$$

$$G.M. = (52 \times 57 \times 50 \times 68 \times 74 \times \dots \times 66)^{1/208} = 59.8147$$

Harmonic Mean (H.M.):

- The 'harmonic mean' is a kind of average of the reciprocals and also the Pythagorean mean.

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

$$H.M. = \frac{208}{\frac{1}{52} + \frac{1}{57} + \frac{1}{50} + \dots + \frac{1}{66}} = 56.25897$$

- Note: It can be always noted that **H.M. ≤ G.M. ≤ A.M.**

Quartiles:

- A 'quartile' is a type of quantile, which divides the number of data points into 4 parts of equal lengths, that are known as 'quarters'.

- The 'first quartile' (Q₁) is a middle number between the lowest number and the median of the data set. Hence in an ordered dataset, Q₁ stands at the boundary between first 25%tile and the rest 75%tile of the data.

- In the given dataset, **Q₁ = 45**, i.e. 25% of the data (208/4 = 52 observations) lie on or on the left of this number and the rest lie on the right side of this number.

- The 'second quartile' (Q₂) is defined as the point, beyond which just 50% of the ordered observations exist. It is also known as the 'median' of the data.

- In the given dataset, **Q₂ = 62**, i.e. 50% of the data (208/2 = 104 observations) lie on or on the left of the dataset.

- The 'third quartile' (Q₃) is defined as the middle number between the largest number and the median of the data set. Hence in an ordered dataset, Q₃ stands at the boundary between first 75%tile and the rest 25%tile of the data.

- In the given dataset, **Q₃ = 83**, i.e. 75% of the data (208×75% = 156 observations) lie on or on the left of this number and the rest lie on the right side of this number.

- The 'fourth quartile' (Q₄) is defined as the largest observed data value, beyond which no more values of the ordered observations exist. It is also known as the 'largest observed value' of the data.

- In the given dataset, **Q₄ = 98**, i.e. 100% of the data (208×100% = 208 observations) lie on or on the left of the dataset.

RESULTS	
Number of Employees	208
Arithmetic Mean (μ)	63.351
Median	62
Mode	53
1st Quartile	45
Median (2nd Quartile)	62
3rd Quartile	83
4th Quartile	98
Sample Variance	428.886
Sample Standard Deviation	20.7096
Population Variance	426.824
Population Standard Deviation	20.6597
Geometric Mean	59.8147
Harmonic Mean	56.259

c. Calculate the variance and the standard deviation.

Variance (σ^2):

- ‘Variance’ measures how far the dataset values are spread out, from their average value. It is the fact or quality of being inconsistent, divergent or different.
- Lower the variance means, lower (narrower) is spread of the data and it is more tightly clumped around a certain mean value.
- Similarly, high variance value indicates a loose widespread (broader) of the data, around a certain mean value.
- Variance (σ^2) is always a non- negative number. $\sigma^2 \geq 0$.
- ‘Population variance’ refers to the value of variance that is calculated from the complete population data (with all the number of samples).
- ‘Sample variance’ is the variance calculated from sample data (with re-scaling to n-1 samples in the denominator).
- Complete steps to calculate variance have been shown in the Excel sheet attached.
- Formulae for Sample variance (σ_s^2) and Population variance (σ_p^2) are given by,

$$\sigma_s^2 = \frac{\sum(x-\mu)^2}{n-1} \quad \text{and} \quad \sigma_p^2 = \frac{\sum(x-\mu)^2}{n} = E[X^2] - \mu^2$$

where, μ = average value or arithmetic mean

n = Number of observations

x = the value of the one observation at a time.

E = Expected value.

- In the given dataset,

$$\sigma_s^2 = 428.8859 \quad \text{and} \quad \sigma_p^2 = 426.824$$

- Note: Since, $1/n < 1/(n-1)$,
hence $\sigma_p^2 < \sigma_s^2$.

- If the observed data values are in Kgs, the variance will be in Kgs², i.e. having the squared unit as that of the original data.

Standard Deviation (σ):

- The Standard Deviation is also a measure of how spread out numbers are, similar to the variance.
- It actually is the positive square root of the variance.
- σ is always a non-negative number. $\sigma \geq 0$.
- Low value indicates low (narrow) spread of the data around a mean value and vice-versa.
- Formulae for Sample Standard Deviation (σ_s) and Population Standard Deviation (σ_p) are given by,

$$\sigma_s = \sqrt{\frac{\sum(x-\mu)^2}{n-1}} \quad \text{and} \quad \sigma_p = \sqrt{\frac{\sum(x-\mu)^2}{n}}$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

where, μ = average value or arithmetic mean

n = Number of observations

x = the value of the one observation at a time

σ_s = Sample Standard Deviation

σ_p = Population or Universal Standard Deviation.

- Complete steps to calculate SD have been shown in the Excel sheet attached.
- In the given dataset,

$$\sigma_s = 20.70956 \quad \text{and} \quad \sigma_p = 20.65972.$$

- Note: Since, $1/n < 1/(n-1)$,
hence $\sigma_p < \sigma_s$.

- If the observed data values are in Kgs, the standard deviation will also be in Kgs, i.e. having the same unit as that of the original data.
- Another important aspect of the SD is the fact that it tells about the most likely range of the data. Most of the given data lies in the range $\mu \pm \sigma_s$, i.e. **between 43 and 84** approximately.

$x - \mu$	$(x - \mu)^2$	Stepwise calculation of Variance and SD		
-11.35096154	128.844328			
-6.350961538	40.3347125	$\Sigma(x - \mu)^2$	88779.38	
-13.35096154	178.248174			
4.649038462	21.6135586	Sample Variance	428.8859	$\Sigma(x - \mu)^2/207$
10.64903846	113.40202	Sample SD	20.70956	$\sqrt{\Sigma(x - \mu)^2/207}$
-31.35096154	982.882789			
-27.35096154	748.075097	Population Variance	426.8239	$\Sigma(x - \mu)^2/208$
-27.35096154	748.075097	Population SD	20.65972	$\sqrt{\Sigma(x - \mu)^2/208}$
27.64903846	764.469328			
33.64903846	1132.25779			
29.64903846	879.065482			
-19.35096154	374.459712			
19.64903846	386.084712			
-8.350961538	69.7385586			
-3.350961538	11.2289432			
-15.35096154	235.65202			
19.64903846	386.084712			
-24.35096154	592.969328			
-9.350961538	87.4404817			

Results in R:

```

> table= read.csv("C:/Users/Downloads/Book2.csv", header= F)
> table2= as.data.frame(as.matrix(c(table[,1],table[,2],table[,3],
      table[,4],table[,5],table[,6],table[,7],table[,8])))

>
> length(unique(table2[,1]))           #67 unique scores
[1] 67
> table(table2[,1])                    #Frequency table

30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
  1  3  2  6  6  2  6  1  4  5  5  1  4  3  2  3  2

47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
  2  4  1  2  3  1  8  5  4  2  2  2  3  3  5  4  5

64 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 82
  4  4  2  4  2  2  2  1  1  2  2  3  5  1  2  3  2

83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98
  4  1  2  4  1  5  4  3  5  5  4  4  5  2  1  4

> sum(table(table2[,1]))               #208 employees
[1] 208

>
> cumsum(table(table2[,1]))            #Cumulative Frequency
30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
  1  4  6 12 18 20 26 27 31 36 41 42 46 49 51 54 56

47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
58 62 63 65 68 69 77 82 86 88 90 92 95 98 103 107 112

64 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 82
116 120 122 126 128 130 132 133 134 136 138 141 146 147 149 152 154

83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98
158 159 161 165 166 171 175 178 183 188 192 196 201 203 204 208

>
>
> mean(table2[,1])                     #Arithmetic Mean (average)
[1] 63.35096

> median(table2[,1])                   #Median
[1] 62
> a= table2[,1];      n= length(a)
> names(table(a))[table(a)==max(table(a))]      #Mode
[1] "53"

> quantile(a)                          #Four Quartiles
  0%  25%  50%  75% 100%
30  45  62  83  98

>
>
> svar= sum((a - mean(a))^2)/(n-1)
> svar                                #Sample Variance
[1] 428.8859
> svar**0.5                           #Sample Standard Deviation
[1] 20.70956

>
> svar*(n-1)/n                         #Population Variance
[1] 426.8239
> (svar*(n-1)/n)**0.5                  #Population Standard Deviation
[1] 20.65972

>
>
> prod(a**(1/n))                       #Geometric mean
[1] 59.81474
> n/(sum(1/a))                         #Harmonic mean
[1] 56.25897
>

```

3. A. In continuation with the data of performance scores of employees in previous example, perform the following:

- Calculate the range and interquartile range.
- Calculate the z- scores.
- Calculate the skewness and Kurtosis (using excel).
- Comment on the distribution of the data.

Range:

- The 'range' of a data is the difference between the amount of highest and lowest values.

$$\text{Range} = \text{Highest value} - \text{Lowest value}$$

$$\text{Range} = 98 - 30 = 68.$$

- Disadvantage— However, the range can be misleading, at giving the idea about values possible to be seen in the data.

- In the given data, other possible values, that are not seen yet, are 0- 29, 99 and 100. Here, the lowest possible value= 0 and the highest possible value= 100. Hence,
maximum possible range= $100 - 0 = 100$.

- For example, {8,11,5,9,1,6,3616} for this set, the range= $3616 - 1 = 3615$.

But, except just one data value, all the other data values are around 10. Hence, IQR and σ are to be calculated.

Inter-Quartile Range (IQR):

- The 'inter-quartile range' (IQR or Midspread) is a measure of variability, on the basis of dividing the dataset into the quartiles.
- IQR is the difference between the amount of largest and smallest values, in the middle 50% of the dataset.

$$\text{Inter-Quartile Range} = \text{Third Quartile} - \text{First Quartile.}$$

$$\text{IQR} = Q_3 - Q_1.$$

But, $Q_1 = 45$ and $Q_3 = 83$. Therefore,

$$\text{IQR} = 83 - 45 = 38.$$

Z- scores (Standardised values):

- A 'z-score' describes the standardised position of data value, as its distance from the mean value.
- A data value has the positive z-score, if it lies above the mean value. Similarly, the z-score is negative, if the data value lies below the mean.
- The z-score (standard score) is the amount of standard deviations, by which a value of the data is below or above the mean (μ).
- Note: The average of all the z-scores must be equal to 0. Sample SD of all the z-scores must be 1.

- Formula for the z-score of a data value (x) is given by:

$$\text{z-score} = \frac{x - \mu}{\sigma}$$

where, σ = sample standard deviation = 20.70956

μ = Arithmetic mean = 63.35096 .

- For example, {1,2,3,4,5} dataset has $\mu = 3$ and $\sigma_s = 1.58$. Hence, z-scores will be: {-1.265, -0.632, 0, 0.632, 1.265}.
- The z-scores have been displayed in the corresponding Excel sheet. For the provided dataset of employees' performances, some of the z- scores are:

Employee Scores (x)	$x - \mu$	$Z = \frac{x - \mu}{\sigma}$	Employee Scores (x)	$x - \mu$	$Z = \frac{x - \mu}{\sigma}$
52	-11.35096	-0.549425209	61	-2.35096	-0.113794489
57	-6.35096	-0.307408142	91	27.64904	1.338307913
50	-13.35096	-0.646232036	50	-13.35096	-0.646232036
68	4.64904	0.225029405	33	-30.35096	-1.469090064
74	10.64904	0.515449886	64	0.64904	0.031415751
32	-31.35096	-1.517493478	39	-24.35096	-1.178669584
36	-27.35096	-1.323879824	88	24.64904	1.193097673
36	-27.35096	-1.323879824	70	6.64904	0.321836232
91	27.64904	1.338307913	33	-30.35096	-1.469090064
97	33.64904	1.628728394	64	0.64904	0.031415751
93	29.64904	1.43511474	54	-9.35096	-0.452618383
44	-19.35096	-0.936652517	34	-29.35096	-1.420686651
83	19.64904	0.951080606	92	28.64904	1.386711327
55	-8.35096	-0.404214969	42	-21.35096	-1.033459344
60	-3.35096	-0.162197902	66	2.64904	0.128222578
48	-15.35096	-0.743038863	53	-10.35096	-0.501021796
83	19.64904	0.951080606	41	-22.35096	-1.081862757
39	-24.35096	-1.178669584	53	-10.35096	-0.501021796
54	-9.35096	-0.452618383	69	5.64904	0.273432818
39	-24.35096	-1.178669584	62	-1.35096	-0.065391075
69	5.64904	0.273432818	59	-4.35096	-0.210601316
95	31.64904	1.531921567	89	25.64904	1.241501087
88	24.64904	1.193097673	61	-2.35096	-0.113794489
			54	-9.35096	-0.452618383

Employee Scores (x)	$x - \mu$	$Z = \frac{x - \mu}{\sigma}$	Employee Scores (x)	$x - \mu$	$Z = \frac{x - \mu}{\sigma}$
34	-29.35096	-1.420686651	88	24.64904	1.193097673
45	-18.35096	-0.888249103	30	-33.35096	-1.614300304
40	-23.35096	-1.13026617	38	-25.35096	-1.227072997
63	-0.35096	-0.016987662	95	31.64904	1.531921567
34	-29.35096	-1.420686651	31	-32.35096	-1.565896891
70	6.64904	0.321836232	60	-3.35096	-0.162197902
54	-9.35096	-0.452618383	75	11.64904	0.563853299
94	30.64904	1.483518154	95	31.64904	1.531921567
93	29.64904	1.43511474	94	30.64904	1.483518154
92	28.64904	1.386711327	63	-0.35096	-0.016987662
31	-32.35096	-1.565896891	48	-15.35096	-0.743038863
63	-0.35096	-0.016987662	76	12.64904	0.612256712
71	7.64904	0.370239645	75	11.64904	0.563853299
64	0.64904	0.031415751	66	2.64904	0.128222578
53	-10.35096	-0.501021796	86	22.64904	1.096290846
51	-12.35096	-0.597828623	67	3.64904	0.176625992
63	-0.35096	-0.016987662	56	-7.35096	-0.355811556
86	22.64904	1.096290846	77	13.64904	0.660660126
36	-27.35096	-1.323879824	33	-30.35096	-1.469090064
38	-25.35096	-1.227072997	76	12.64904	0.612256712
49	-14.35096	-0.69463545	39	-24.35096	-1.178669584
96	32.64904	1.580324981	51	-12.35096	-0.597828623
77	13.64904	0.660660126	33	-30.35096	-1.469090064
36	-27.35096	-1.323879824	34	-29.35096	-1.420686651

Employee Scores (x)	$x - \mu$	$Z = \frac{x - \mu}{\sigma}$	Employee Scores (x)	$x - \mu$	$Z = \frac{x - \mu}{\sigma}$
62	-1.35096	-0.065391075	95	31.64904	1.531921567
45	-18.35096	-0.888249103	85	21.64904	1.047887433
86	22.64904	1.096290846	35	-28.35096	-1.372283237
38	-25.35096	-1.227072997	83	19.64904	0.951080606
91	27.64904	1.338307913	55	-8.35096	-0.404214969
71	7.64904	0.370239645	79	15.64904	0.757466953
61	-2.35096	-0.113794489	80	16.64904	0.805870366
94	30.64904	1.483518154	80	16.64904	0.805870366
74	10.64904	0.515449886	58	-5.35096	-0.259004729
57	-6.35096	-0.307408142	98	34.64904	1.677131807
38	-25.35096	-1.227072997	90	26.64904	1.2899045
59	-4.35096	-0.210601316	61	-2.35096	-0.113794489
82	18.64904	0.902677193	61	-2.35096	-0.113794489
98	34.64904	1.677131807	31	-32.35096	-1.565896891
54	-9.35096	-0.452618383	82	18.64904	0.902677193
92	28.64904	1.386711327	91	27.64904	1.338307913
84	20.64904	0.99948402	48	-15.35096	-0.743038863
89	25.64904	1.241501087	98	34.64904	1.677131807
68	4.64904	0.225029405	55	-8.35096	-0.404214969
36	-27.35096	-1.323879824	78	14.64904	0.709063539
68	4.64904	0.225029405	55	-8.35096	-0.404214969
35	-28.35096	-1.372283237	93	29.64904	1.43511474
42	-21.35096	-1.033459344	79	15.64904	0.757466953
95	31.64904	1.531921567	40	-23.35096	-1.13026617

Employee Scores (x)	$x - \mu$	$Z = \frac{x - \mu}{\sigma}$	Employee Scores (x)	$x - \mu$	$Z = \frac{x - \mu}{\sigma}$
96	32.64904	1.580324981	92	28.64904	1.386711327
56	-7.35096	-0.355811556	43	-20.35096	-0.98505593
37	-26.35096	-1.275476411	42	-21.35096	-1.033459344
32	-31.35096	-1.517493478	93	29.64904	1.43511474
34	-29.35096	-1.420686651	83	19.64904	0.951080606
34	-29.35096	-1.420686651	40	-23.35096	-1.13026617
33	-30.35096	-1.469090064	39	-24.35096	-1.178669584
77	13.64904	0.660660126	86	22.64904	1.096290846
90	26.64904	1.2899045	77	13.64904	0.660660126
43	-20.35096	-0.98505593	44	-19.35096	-0.936652517
53	-10.35096	-0.501021796	58	-5.35096	-0.259004729
53	-10.35096	-0.501021796	89	25.64904	1.241501087
98	34.64904	1.677131807	88	24.64904	1.193097673
47	-16.35096	-0.791442277	53	-10.35096	-0.501021796
89	25.64904	1.241501087	87	23.64904	1.14469426
88	24.64904	1.193097673	60	-3.35096	-0.162197902
72	8.64904	0.418643059	48	-15.35096	-0.743038863
80	16.64904	0.805870366	46	-17.35096	-0.83984569
62	-1.35096	-0.065391075	73	9.64904	0.467046472
91	27.64904	1.338307913	68	4.64904	0.225029405
33	-30.35096	-1.469090064	53	-10.35096	-0.501021796
47	-16.35096	-0.791442277			
36	-27.35096	-1.323879824			
51	-12.35096	-0.597828623			

=(A209 - 63.35096)/20.70956

	A	B	C	D
	Employee Scores (x)	$x - \mu$	$Z = \frac{x - \mu}{\sigma}$	
190	42	-21.351	-1.030971204	
191	43	-20.351	-0.982684325	
192	62	-1.35096	-0.065233641	
193	67	3.64904	0.17620075	
194	66	2.64904	0.127913872	
195	76	12.64904	0.610782653	
196	46	-17.351	-0.837823691	
197	77	13.64904	0.659069531	
198	63	-0.35096	-0.016946763	
199	64	0.64904	0.031340115	
200	85	21.64904	1.045364556	
201	92	28.64904	1.383372703	
202	90	26.64904	1.286798947	
203	40	-23.351	-1.12754496	
204	59	-4.35096	-0.210094275	
205	40	-23.351	-1.12754496	
206	53	-10.351	-0.499815544	
207	94	30.64904	1.47994646	
208	45	-18.351	-0.886110569	
209	66	2.64904	0.127913872	
210				
211				
212	63.35096	0	0	Mean
213	20.70956	20.70956	1	Sample SD

Skewness:

- ‘Skewness’ refers to the amount of distortion, in a normal bell curve. Skewness represents the extent, to which a given dataset varies.
- A normal distribution (bell) generally has a skewness of zero.
- A negative skewness indicates that the *tail* of the distribution is on the left side. A positive skewness indicates, the tail to be on the right.
- Application- Investors note skewness, while judging a return distribution because it considers the extremes (outliers) of the dataset instead of focusing only on the means. Kurtosis also does the same and hence used as an alternative.
- Some formulae for measuring skewness are:

$$\text{Pearson's Mode Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$S_1 = \frac{\mu - \text{Mode}}{\sigma} = \frac{63.35096 - 53}{20.70956}$$

$$S_1 = 0.5$$

$$\text{Pearson's Median Skewness} = \frac{3 (\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$S_2 = \frac{3 (\mu - \text{Median})}{\sigma} = \frac{3 (63.35096 - 62)}{20.70956}$$

$$S_2 = 0.2$$

$$\text{Co-efficient of Skewness} = \frac{\Sigma(x - \mu)^3}{(n-1) \cdot \sigma^3} = \frac{\Sigma(x - 63.35096)^3}{207 * 20.70956^3}$$

$$\text{Co-efficient of Skewness} = \mathbf{0.09767} \quad (\text{using MS Excel}).$$

Kurtosis:

- Kurtosis tells us about the height and the sharpness of the central peak, when compared to that of a standard bell curve.
- ‘Kurtosis’ defines, how heavily/ mildly the tails of the data distribution differ from that of a bell distribution. Kurtosis tells, whether the tails of the given data distribution contain extreme values and the number of extreme values.
- Kurtosis is a measure of the ‘tailedness’ (outliers/ extremes) in a probability distribution.
- Formula for measuring kurtosis is:

$$\text{Kurtosis} = \frac{\Sigma(x - \mu)^4}{(n-1) \cdot \sigma^4} - 3 = \frac{\Sigma(x - 63.35096)^4}{207 * 20.70956^4} - 3$$

$$\text{Kurtosis} = \mathbf{-1.2854} \quad (\text{using MS Excel}).$$

RESULTS					
Mean (μ)	63.351		Range	68	
Median	62		IQR	38	
Mode	53				
Population Variance	426.824		SKEWNESS	0.09767	⇐ Approximately Symmetric
Population SD (σ)	20.6597		KURTOSIS	-1.2854	⇐ PlatyKurtic Distribution
Highest value	98				
Lowest value	30	Mode Skewness	0.49982		
1st Quartile (Q1)	45	Median Skewness	0.1957		
2nd Quartile (Median)	62				
3rd Quartile (Q3)	83				

Data distribution:

- Mean= 63.35096, Median= 62, Mode= 53.

Median < Mean and Mode < Mean.

Hence, it appears that the data may be positively skewed (towards the right).

- If Skewness $\notin (-1,1)$; highly skewed distribution.

If Skewness $\in (-1, -0.5)$ or Skewness $\in (0.5, 1)$; moderately skewed distribution.

If Skewness $\in [-0.5, 0.5]$; approximately symmetric distribution.

- However, Co-efficient of Skewness = 0.097.

Hence, the distribution is **approximately symmetric**.

- If Kurtosis ≈ 0 ; Mesokurtic distribution $\Rightarrow$ Close to Normal distribution.

If Kurtosis > 0 ; Leptokurtic distribution $\Rightarrow$ More outliers, heavy tails, risky financial investments.

If Kurtosis < 0 ; Platykurtic distribution $\Rightarrow$ **Less outliers, low tails**, desirable financial investments.

- Since, Kurtosis= -1.2854. Hence, the data follows a **Platykurtic distribution**.

- First Quartile= 45 and Third Quartile= 83. IQR= 38.

• Hence, on the performance score's scale between 0 and 100, the middle 50% of the data lies between just 45 and 83, i.e. just **38% (nearly 3/8th) of all possible performance scores**.

- Range = 68.

Out of all the possible scores between 0 and 100, every employee is having the score in 68% range (nearly 2/3rd).

About 1/3rd of the scores are not obtained by anybody.

Results in R:

```
> (mean(a)- 53)/ sd(a)           #Pearson's Mode Skewness
[1] 0.4998156
> 3*(mean(a)- median(a))/ sd(a)  #Pearson's Median Skewness
[1] 0.1957011
>
> max(a)- min(a)                 #Range
[1] 68
> as.numeric(quantile(a)[4]- quantile(a)[2]) #IQR
[1] 38
>
> sum((a- mean(a))**3)/ ((n-1)*(sd(a)**3))    #Skewness
[1] 0.09673192
> sum((a- mean(a))**4)/ ((n-1)*(sd(a)**4)) -3  #Kurtosis
[1] -1.29169
> |
```

$(x - \mu)^3$	$(x - \mu)^4$	Stepwise calculation of co-efficients of Skewness and Kurtosis		
-1462.506	16600.852			
-256.164	1626.8875	$\Sigma(x - \mu)^3$	177850	
-2379.784	31772.397	$[\Sigma(x - \mu)^3] / (n-1)$	859.1786	
100.48237	467.14653	$[\Sigma(x - \mu)^3] / [(n-1) \cdot \sigma^3]$	0.096732	← SKEWNESS
1207.623	12860.026			
-30814.32	966058.39	$\Sigma(x - \mu)^4$	65046006	
-20460.57	559616.22	$[\Sigma(x - \mu)^4] / (n-1)$	314231.9	
-20460.57	559616.22	$[\Sigma(x - \mu)^4] / [(n-1) \cdot \sigma^4]$	1.70831	
21136.845	584413.48	$[\Sigma(x - \mu)^4] / [(n-1) \cdot \sigma^4] - 3$	-1.29169	← KURTOSIS
38099.391	1282007.9			
26063.45	772756.28			
-7246.154	140220.03			
7586.1951	149061.45			
-582.3837	4863.463			
-37.62771	126.08893			
-3617.484	55531.852			
7586.1951	149061.45			
-14439.37	351612.53			
-817.6522	7645.8328			
-14439.37	351612.53			
180.2702	1018.3536			
31701.632	1003326.2			
14976.145	369147.59			

3. B. In continuation with the data of performance scores of employees in previous example, perform the following:

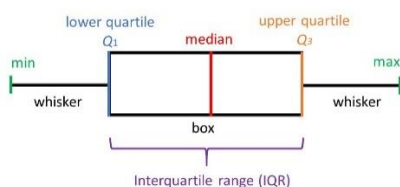
- a. Make the histogram
- b. Plot the box-plot diagram
- c. Plot the frequency polygon
- d. Plot the Ogive diagram.

Histogram:

A ‘histogram’ is a graphical plot/ display of the given data, using bars of different frequencies (heights). Taller bars indicate more data falling into the particular range. A histogram also displays the spread and shape of the continuous given sampled data.

Box-Plot:

A box- plot graphically depicts groups of given data, with the help of their quartiles. Box- plots also have extended lines, indicating variability outside the first and third quartiles. Hence, it is also known as the terms ‘box-and-whisker plot’ diagram. It contains minima, Q1, median, Q3 and maxima in this proper sequence only.



Frequency Polygon:

A ‘frequency polygon’ graph is constructed by using histogram bars/ bins/ intervals. There are lines to join midpoints of each bin. Here, the heights of bins are converted into points, that represent frequencies. A frequency polygon is generally created from the histogram. From the frequency distribution table, it can also be created by calculating midpoints of the intervals.

Ogive Diagram:

An 'ogive diagram' (cumulative frequency polygon) shows cumulative frequencies. Here, the cumulative percentages are added on graph from left to right (lower range to higher range). An ogive graph has cumulative frequencies on the y-axis and class ranges (boundaries) on the x-axis.

Conclusion:

- The histogram does **not** follow any normal distribution.
- However, the ogive diagram suggests that the data bins are more likely to be **flat**, rather than curved, i.e. **nearly equal number of employees' fall under each bin**.
- However, middle 50% of distribution has scores between 45 and 83 (Q1 and Q3).
- On the scale of 0 to 100, the 1/3rd of the scores have not been obtained by anybody.
- The data is approximately symmetric.

